UNIVERSITY ABDELMALEK ESSAADI
National School of Sciences
Applied from Al Hoceima

detection of fraud in online transactions

USING MACHINE LEARNINIG MODEL

realized by
WANAIM essaadia

supperviser by
Pr.khamjane

Fraud Detection using Machine Learning

Abstract— Recent research has shown that machine learning techniques have been applied very effectively to the problem of payments related fraud detection. Such ML based techniques have the potential to evolve and detect previously unseen patterns of fraud. In this paper, we apply multiple ML techniques based on Logistic regression and Random Forest Model to the problem of payments fraud detection using a labeled dataset containing payment transactions. We show that our proposed approaches are able to detect fraud transactions with high accuracy and reasonably low number of false positives.

INTRODUCTION

We are living in a world which is rapidly adopting digital payments systems. Credit card and payments companies are experiencing a very rapid growth in their transaction volume. In third quarter of 2018, PayPal Inc (a San Jose based payments company) processed 143 billion USD in total payment volume [4]. Along with this transformation, there is also a rapid increase in financial fraud that happens in these payment systems.

An effective fraud detection system should be able to detect fraudulent transactions with high accuracy and efficiency. While it is necessary to prevent bad actors from executing fraudulent transactions, it is also very critical to ensure genuine users are not prevented from accessing the payments system. A large number of false positives may translate into bad customer experience and may lead customers to take their business elsewhere.

A major challenge in applying ML to fraud detection is presence of highly imbalanced data sets. In many available datasets, majority of transactions are genuine with an extremely small percentage of fraudulent ones. Designing an accurate and efficient fraud detection system that is low on false positives but detects fraudulent activity effectively is a significant challenge for researchers.

n our paper, we apply multiple binary classification approaches - Logistic regression , Random Forest and XGBoost on a labeled dataset that consists of payment transactions.

Our goal is to build binary classifiers which are able to separate fraud transactions from non-fraud transactions. We compare the effectiveness of these approaches in detecting fraud transactions.

RELEVANT RESEARCH

Several ML and non-ML based approaches have been applied to the problem of payments fraud detection. The paper [1] reviews and compares such multiple state of the art techniques, datasets and evaluation criteria applied to this problem. It discusses both supervised and unsupervised ML based approaches involving ANN (Artificial Neural Networks), SVM (Support Vector machines) ,HMM (Hidden Markov Models), clustering etc. The paper [5] proposes a rule based technique applied to fraud detection problem. The paper [3] discusses the problem of imbalanced data that result in a very high number of false positives and proposes techniques to alleviate this problem. In [2], the authors propose an SVM based technique to detect metamorphic malware. This paper also discusses the problem of imbalanced data sets - fewer malware samples compared to benign files - and how to successfully detect them with high precision and accuracy.

Approach

- **Exploratory Data Analysis (EDA):** It involves the use of statistical and graphical techniques to explore the characteristics of a dataset, understand its underlying structure, and uncover patterns, trends, relationships, and potential outliers.
- **Data Preprocessing:** Address any missing or anomalous values in the dataset. Encode categorical variables, if necessary. Scale numerical features to ensure uniformity in their impact on the model.
- **Handling Imbalance:** Implement techniques to address class imbalance, such as oversampling the minority class (fraudulent transactions) using methods like SMOTE or undersampling the majority class.
- **Model Selection:** Experiment with different classification algorithms such as Logistic Regression, Random Forest, XGBoost, etc. Utilize appropriate evaluation metrics for imbalanced datasets, emphasizing metrics like precision, recall, F1-score, and ROC-AUC.
- **Hyperparameter Tuning:** Conduct a systematic search for optimal hyperparameters using techniques like GridSearchCV or RandomizedSearchCV to improve model performance.
- **Model Evaluation:** Evaluate the performance of the trained model on a separate test dataset. Assess the model's ability to correctly classify fraudulent transactions while minimizing false positives.
- **Interpretability and Explainability:** Strive for a model that not only performs well but is also interpretable. Understand the importance of each feature in the decision-making process.

DATASET AND ANALYSIS

In this project, we have used a Kaggle provided dataset [8] of simulated mobile based payment transactions. We analyze this data by categorizing it with respect to different types of transactions it contains.
The dataset contains five distinct types of transactions in the dataset: PAYMENT, TRANSFER, CASH_OUT, DEBIT, and CASH_IN.
The distribution of these transaction types is as follows:

- CASH_OUT: 2,237,500 transactions
- PAYMENT: 2,151,495 transactions
- CASH_IN: 1,399,284 transactions
- TRANSFER: 532,909 transactions
- DEBIT: 41,432 transactions

the dataset consists of both numerical and categorical features like transaction type ,amount transferred, account numbers of sender and recipient accounts. In our experiments we use the following features to train our models.
1) Transaction type
2) Transaction amount
3) Sender account balance before transaction
4) Sender account balance after transaction
5) Recipient account balance before transaction
6) Recipient account balance after transaction
The dataset consists of around 6 million transactions out of which 8312 transactions are labeled as fraud. It is highly imbalanced with 0.13 percent fraud transactions.(figure 1)

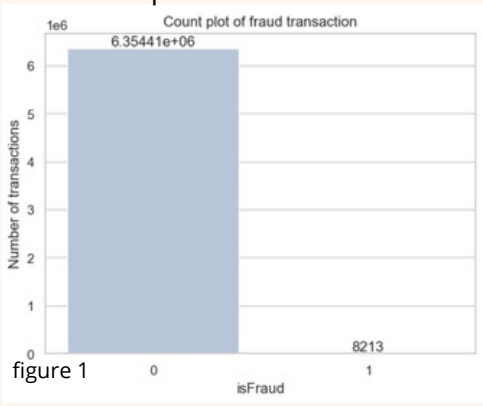


figure 1

the oncentration of fraud within CASH_OUT and TRANSFER transactions suggests that these specific types may be more susceptible to fraudulent activities.(figure 2)

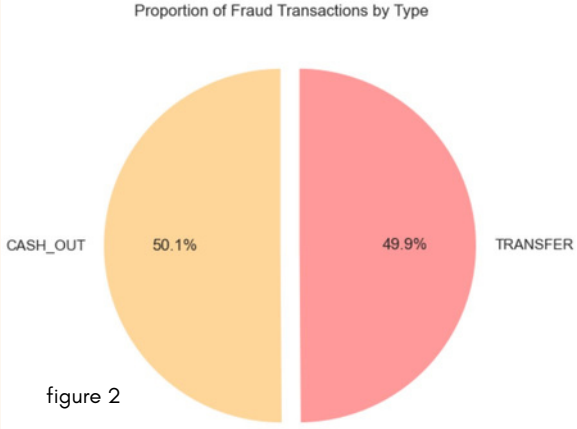


figure 2

Goal
The primary objective of this analysis is to develop a robust classification model that can accurately predict whether a given online transaction is fraudulent or not. Leveraging the features provided in the dataset, the goal is to train a machine learning model capable of distinguishing between legitimate and fraudulent transactions.

Statistics on the amounts for non-fraud transactions:
count 6,354,407.00
mean 178,197.04
std 596,236.98
min 0.01
25% 13,368.40
50% 74,684.72
75% 208,364.76
max 92,445,516.64
Name: amount, dtype: float64
Statistics on the amounts for fraud transactions:
count 8,213.00
mean 1,467,967.30
std 2,404,252.95
min 0.00
25% 127,091.33
50% 441,423.44
75% 1,517,771.48
max 10,000,000.00
Name: amount, dtype: float64

Key Observations:

- Fraudulent transactions, on average, involve significantly higher amounts compared to non-fraudulent transactions.
- The standard deviation for fraud transactions is notably higher, indicating a wider range of transaction amounts.
- The minimum amount for fraud transactions is $0, suggesting instances of negligible or anomalous values in fraudulent activities.
- The upper percentiles (75th and maximum) for fraud transactions demonstrate a substantial increase in the transaction amounts, further emphasizing the contrast between the two categories.

Many fraud transactions are linked to customers whose account balances drop to zero afterward. This pattern implies that fraudsters often manipulate transactions to deplete or minimize the affected customers' balances. Detecting and keeping an eye on instances where substantial transactions result in zero balances is crucial for preventing and identifying fraud. Understanding this pattern helps improve security measures and safeguards against fraudulent activities that aim to drain customer accounts.(figure 3)

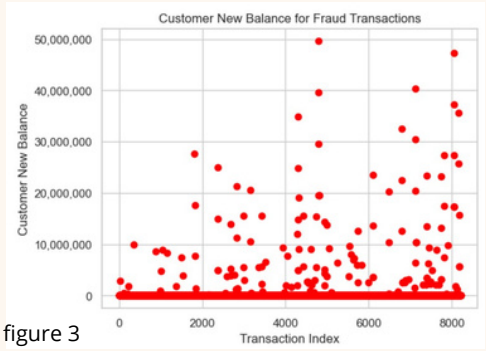


figure 3

While the heatmap may not reveal strong correlations with the target variable (isFraud), it does highlight a noteworthy observation. There is a significant correlation between the new and old balances of accounts. Although the heatmap may not directly assist in understanding fraud patterns, recognizing the correlation between certain features, such as old and new balances, can be valuable for refining feature engineering and developing more nuanced models for fraud detection(figure 4).

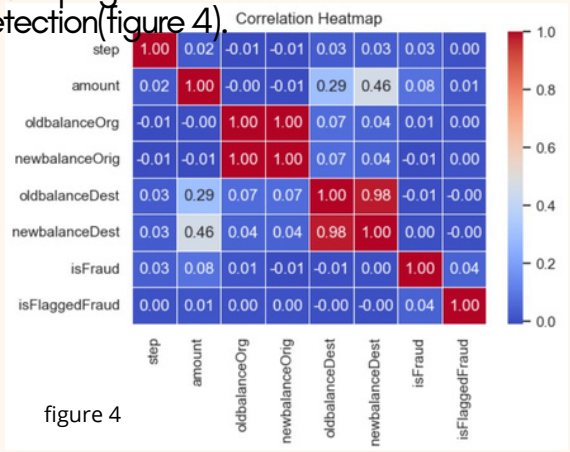
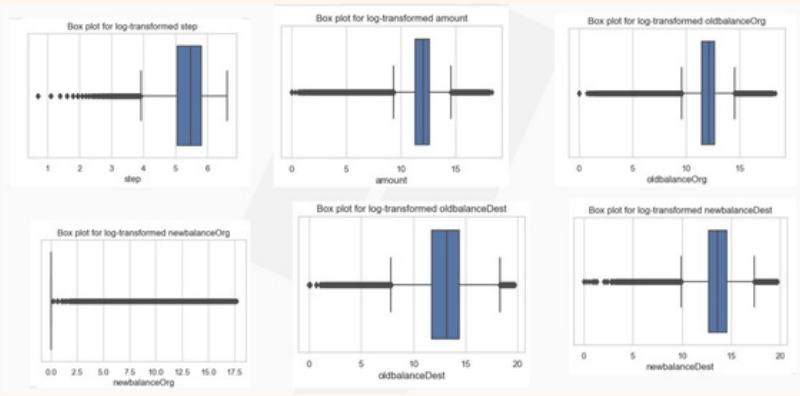


figure 4

To optimize the dataset and streamline it for analysis, transactions of types 'CASH_IN', 'DEBIT', and 'PAYMENT' have been excluded from the process_df DataFrame. This decision is grounded in the observation that there are no instances of fraud associated with these transaction types. By excluding these types, we aim to minimize the dataset proportions and focus on transaction categories that have relevance to the identification and analysis of fraudulent activities, thus enhancing the efficiency of subsequent modeling and analysis efforts.

We've substantially reduced the dataset, now containing 2,770,409 entries converting categorical of the 'type' column data into a format suitable for machine learning models

The 'step,' 'amount,' 'oldbalanceOrg,' 'newbalanceOrg,' 'oldbalanceDest,' and 'newbalanceDest' columns were log-transformed using np.log1p() to make the distributions more symmetric.
The output showcases the log-transformed values for each column, and the subsequent box plots visually represent the distribution of these log-transformed values. It's important to note that the box plots are based on log-transformed values for better visualization and outlier detection.
The presence of numerous points outside the box in the box plots indicates the potential existence of outliers or extreme values. These outliers may significantly deviate from the majority of the data and could impact the statistical analysis or modeling processes. figure5



The RobustScaler is employed to scale features based on the interquartile range (IQR), ensuring robustness to outliers. This scaler is particularly beneficial when dealing with data that contains outliers, as it leverages the median and IQR for scaling rather than the mean and standard deviation.
In this specific scenario, the log1p transformation is applied to the data, addressing cases where the IQR is 0, thus enabling a more robust scaling mechanism. The combination of log1p and RobustScaler contributes to the preprocessing steps, enhancing the distribution of data and providing a resilient scaling approach suitable for features with potential outliers.

Evaluating different models:

We are splitting the data into training (80%) and testing (20%) sets. To ensure that the class distribution is preserved in both the training and testing sets, we use the stratify parameter. This is particularly important when dealing with imbalanced datasets where the distribution of classes is uneven. Here's how you can implement it

1) Logistic regression

Logistic regression performs the binary classification by using a sigmoid function as the hypothesis, which is given by:

$$P(y = 1|x; \theta) = h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

The logistic regression model is trained by fitting the parameter θ via maximum likelihood, where the log likelihood function can be represented as:

$$\ell(\theta) = \sum_{i=1}^{n} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log\left(1 - h(x^{(i)})\right)$$

Then, θ can be updated using stochastic gradient ascent rule

$$\theta_j := \theta_j + \alpha \frac{\partial}{\partial \theta_j} \ell(\theta)$$

$$:= \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

Random forest classifier

Tree classification is very powerful to classify the nonlinear dataset, like NLP. The classification includes bagged tree, random forest, and boosting [8]. Random forest provides an improvement over the bagged trees. Bagged trees consider all the predictors (p predictors) in every split of the tree, whereas random forest limits the selection of the predictors to m predictors. The number of predictors considered in the split in random forest is equal to the square root of the total number of predictors, $m = \sqrt{p}$. In other words, random forest decorrelates the trees through considering less predictors. Unlike highly correlated bagged trees, the variance in random forest is significantly decreased [8].

The setting of random forest in this report:

- The number of trees: 100
- Quality criterion: Gini index.
  K is the class number. M is the sample size. The value will take on a small value if the node is pure.

$$G = \sum_{K=1}^{K} p_{mk} \, log \, p_{mk}$$

- The maximum depth of the tree: None
- The minimum number of samples required to split an internal node: 2

The matrix provides several evaluation parameters, including:

- Positive precision: the accuracy of the positive prediction.

$$Positive\ precision = \frac{TP}{TP + FP}$$

- Negative precision: the accuracy of the negative prediction.

$$Negative\ precision = \frac{TN}{TN + FN}$$

- Accuracy: the ratio of the correct predictions, which is the average of negative and positive precision.

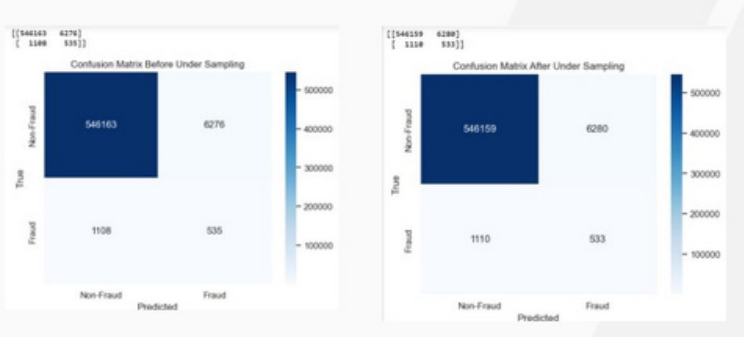
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

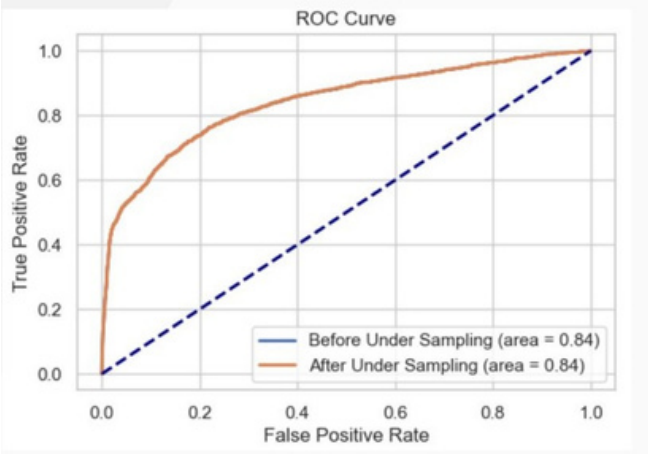
The performance evaluation of the Logistic Regression models on both oversampled and undersampled datasets reveals similar results.
For the Oversampled Dataset:

- Precision: The model achieves a high precision of 1.00 for the majority class (non-fraud), indicating a low false positive rate.
- Recall: The recall for the minority class (fraud) is 0.33, suggesting that the model captures a third of the actual fraud cases.
- F1-Score: The F1-score for fraud detection is 0.13, reflecting a trade-off between precision and recall.
- Matthews Correlation Coefficient (MCC): The MCC is 0.16, indicating a moderate correlation between predicted and actual classes.
- ROC-AUC Score: The ROC-AUC score is 0.84, representing a good balance between true positive rate and false positive rate.

For the Undersampled Dataset:

- Precision, Recall, and F1-Score: The precision, recall, and F1-score metrics for both classes are identical to the oversampled case, reflecting similar classification performance.
- Matthews Correlation Coefficient (MCC): The MCC is 0.15, showing a comparable correlation with the actual classes.
- ROC-AUC Score: The ROC-AUC score is also 0.84, aligning with the performance on the oversampled data.





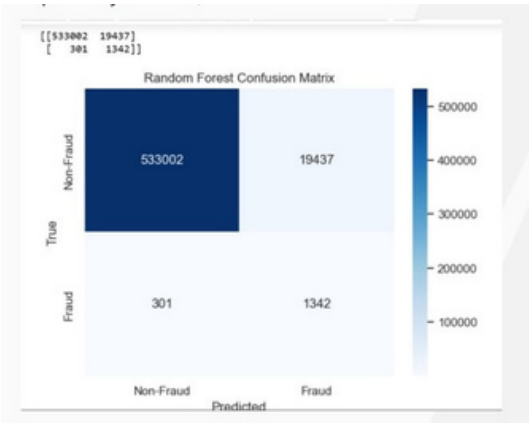
Analysis:

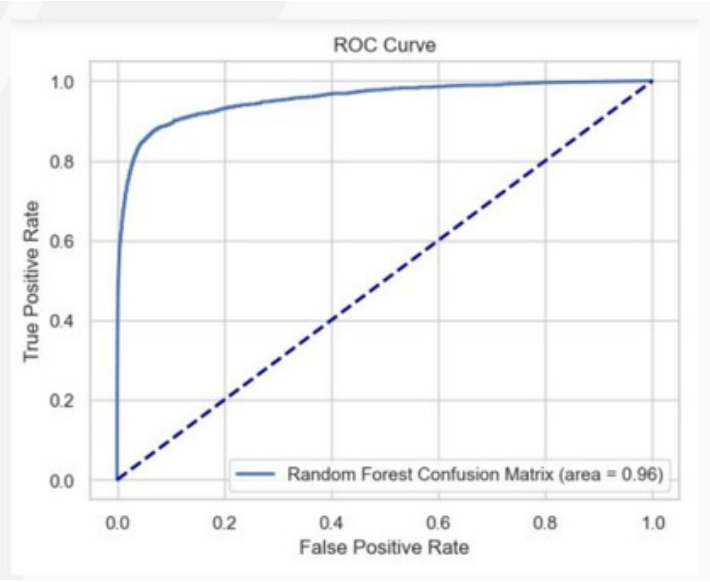
- True Positives (Fraud Detected): The number of true positive cases remained relatively stable, decreasing by only 2 from 535 to 533 after undersampling.
- True Negatives (Non-Fraud Detected): The model maintained a high true negative count, indicating a consistent ability to correctly identify non-fraudulent transactions.
- False Positives (False Alarms): The number of false positives increased slightly from 6,276 to 6,280. While this might suggest a slight reduction in precision, the overall impact is not substantial.
- False Negatives (Missed Fraud): The number of false negatives also increased slightly from 1,108 to 1,110. This indicates a marginal decrease in recall, suggesting that a few more fraudulent transactions were missed after undersampling.

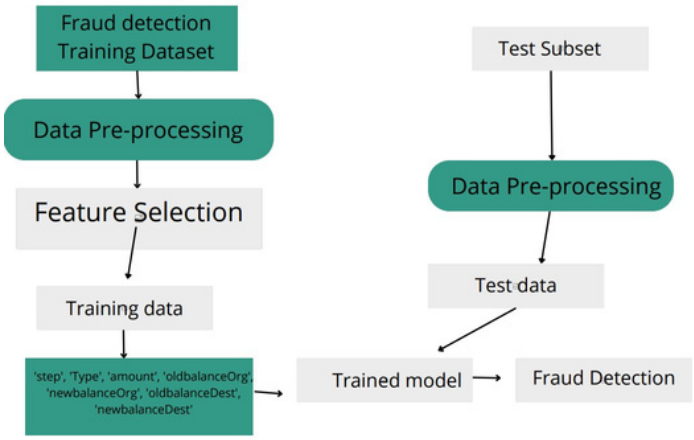
Conclusion:

- The model's performance, as indicated by the confusion matrices, remains relatively stable before and after undersampling.
- The decrease in recall after undersampling suggests a slight reduction in the model's ability to identify all instances of fraud. However, this needs to be considered in the context of the trade-off between precision and recall.

Random Forest
For highly unbalanced datasets, an ensemble method like Random Forest can often perform well. Random Forest is an ensemble of decision trees, and its ability to combine multiple trees can help mitigate the impact of class imbalance.







Comparing the Random Forest model with the logistic regression model, both trained on the under-sampled dataset, we observe distinct performance characteristics.
Random Forest Model:
Advantages:

- Achieves a higher Matthews Correlation Coefficient (MCC) of 0.27, indicating a stronger overall performance.
- Demonstrates a robust ROC-AUC score of 0.95, reflecting a high ability to distinguish between classes.
- Shows notable recall for the minority class (fraudulent transactions) at 0.76, highlighting its effectiveness in capturing positive instances.

Considerations:

- Precision for the minority class is low (0.10), indicating a higher rate of false positives.
- Logistic Regression Model (Under-Sampled):
- Advantages:
- Higher precision for the minority class at 0.08, implying fewer false positives compared to the Random Forest model.
- Maintains a respectable ROC-AUC score of 0.84.

Considerations:

- MCC is lower at 0.15, suggesting a moderate overall performance. Recall for the minority class is lower at 0.32, indicating that it captures fewer positive instances compared to the Random Forest model.

- The Random Forest model has a higher True Positive count, indicating a better ability to correctly identify fraudulent transactions.
- However, it also has a higher False Positive count, suggesting a greater tendency to classify non-fraudulent transactions as fraudulent.
- The Logistic Regression model, while having fewer True Positives, also exhibits fewer False Positives.
- The choice between the two models depends on the specific requirements of the application, considering the trade-off between correctly identifying fraud and minimizing false alarms.

REFERENCES :

- [1] A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective - Samaneh Sorournejad, Zojah, Atani et.al - November 2016 [2] Support Vector machines and malware detection - T.Singh,F.Di Troia, C.Vissagio , Mark Stamp - San Jose State University - October 2015 [3] Solving the False positives problem in fraud prediction using automated feature engineering - Wedge, Canter, Rubio et.al - October 2017 [4] PayPal Inc. Quarterly results https://www.paypal.com/stories/us/paypalreports-third-quarter-2018-results [5] A Model for Rule Based Fraud Detection in Telecommunications - Rajani, Padmavathamma - IJERT - 2012 [6] HTTP Attack detection using n-gram analysis - A. Oza, R.Low, M.Stamp - Computers and Security Journal - September 2014 [7] Scikit learn - machine learning library http://scikit-learn.org [8] Paysim - Synthetic Financial Datasets Fo